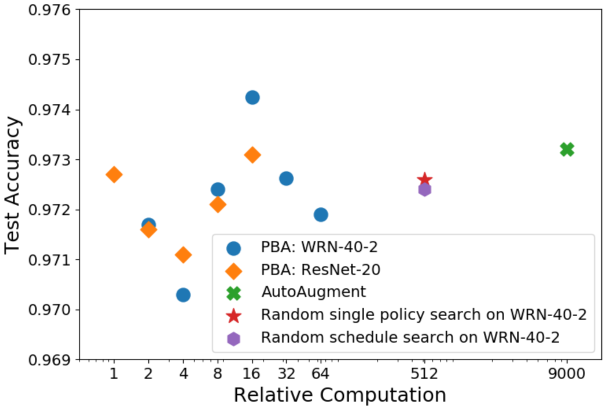


Figure 1. Comparison of relative cost of search to test accuracy of augmentation policies evaluated on WideResNet-28-10. Child model WRN-40-2 was evaluated for population sizes from 2 to 64, and ResNet-20 was evaluated for sizes 2 to 32. All policies were trained on Reduced CIFAR-10.

A. PBA Scalability with Compute

In Figure 1 we look at how large the model and PBT population size is necessary to learn an effective schedule. The population size determines how much of the search space is explored during training, and also the computational overhead of PBA. Our results indicate that a population size of 16 WRN-40-2 models performs the best. Having more than 16 trials seems not to help, and having less than 16 seems to lead to decreased performance. However, we found that results could fluctuate significantly between runs of PBT, most likely due to exploring a very limited search space with a noisy exploration strategy.

Besides WRN-40-2, we also tried to use a ResNet-20 (He et al., 2016) model for PBT population, which required about half the compute. Empirical results (in Table 1 and Figure 1) suggest that the ResNet-20 population does not achieve as high of a test accuracy as with WRN-40-2, but results were relatively close. Because a ResNet-20 model has much less parameters, training accuracy plateaus faster than WRN-40-2, which may change the effects of augmentation.

B. Model Hyperparameters

The hyperparameters used to train WideResNet-40-2 to discover augmentation schedules, and also the ones used to train final models, are displayed in Table 2. For full details on the hyperparameters and implementation, see the open source code.

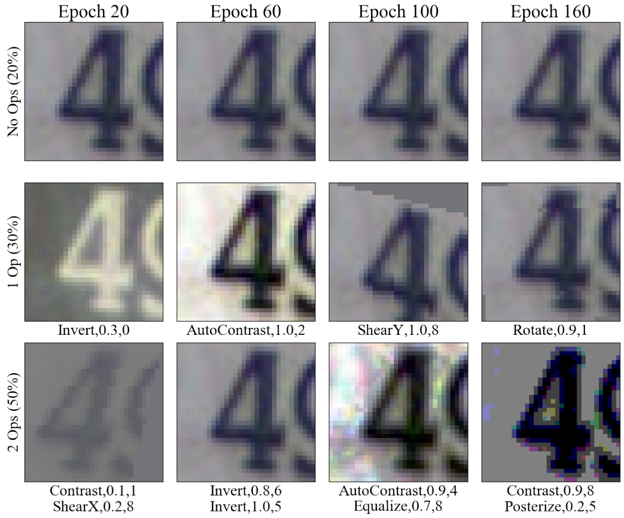


Figure 2. Augmentations applied to a SVHN "4" class image, at various points in our augmentation schedule learned on Reduced SVHN data. Each operation is formatted with name, probability, and magnitude value respectively.

C. SVHN discovered schedule

See Figure 2 for a visualization of the policy on an example image and Figure 3 for a visualization of an example PBA policy on the SVHN dataset.

Examining the learned policy schedule, we observe that Cutout, Translate Y, Shear X, and Invert stand out as being present with high probability across all epochs. This fits with the findings of (Cubuk et al., 2018) indicating that Invert and geometric transformations are successful in SVHN because it is important to learn invariances to these augmentations. From another perspective, all of the augmentations appear with reasonable probability at some point in the schedule, which suggests that using a preliminary strategy like AutoAugment to filter out poor performing augmentations would be an interesting direction to explore.

References

Cubuk, E. D., Zoph, B., Mané, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation policies from data. *CoRR*, abs/1805.09501, 2018. URL http://arxiv.org/abs/1805.09501.

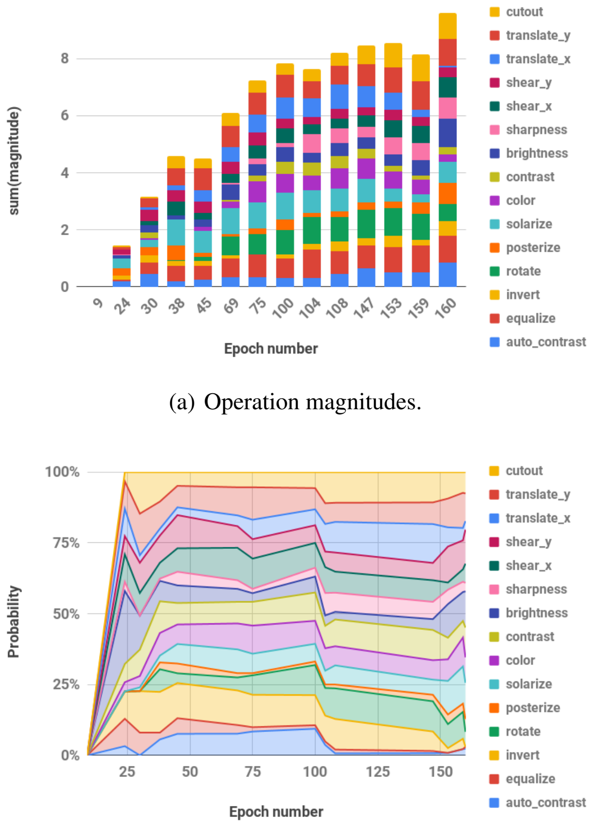
He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. *CoRR*, abs/1603.05027, 2016. URL http://arxiv.org/abs/1603.05027.

Table 1. Test error during PBT search and policy schedule evaluated afterwards, for varying population sizes and models. PBA Search with variation of model and compute, on Reduced CIFAR-10 dataset. ResNet-20 (Res) took approximately half the compute of WideResNet-40-2 (WRN). Number in title is the population size, and speedup is relative to AutoAugment. Note that models with larger population sizes, while scoring high during the search, don't actually perform better when re-evaluated.

Model	8-Res	16-Res	32-Res	16-WRN	32-WRN	64-WRN
WRN-40-2 during search	-	-	-	0.8484	0.8446	0.8523
WRN-40-2	-	-	-	0.8452	0.8445	0.8446
ResNet-20 during search	0.7484	0.7657	0.7619	-	-	-
ResNet-20	0.7457	0.7545	0.7534	-	-	-
WRN-28-10	0.9711	0.9721	0.9740	0.9740	0.9736	0.9703
Relative Speedup	2250x	1125x	562.5x	562.5x	281.25x	140.625x

Table 2. Hyperparameters used for evaluation on CIFAR-10, CIFAR-100, and (R)educed-CIFAR-10. Besides Wide-ResNet-28-10 and Wide-ResNet-40-2 on Reduced SVHN, no hyperparameter tuning was done. Instead, all hyperparameters are the same as those used in AutoAugment.

Dataset	Model	Learning Rate	Weight Decay	Batch Size
CIFAR-10	Wide-ResNet-40-2	0.1	0.0005	128
CIFAR-10	Wide-ResNet-28-10	0.1	0.0005	128
CIFAR-10	Shake-Shake (26 2x32d)	0.01	0.001	128
CIFAR-10	Shake-Shake (26 2x96d)	0.01	0.001	128
CIFAR-10	Shake-Shake (26 2x112d)	0.01	0.001	128
CIFAR-10	PyramidNet+ShakeDrop	0.05	0.00005	64
CIFAR-100	Wide-ResNet-28-10	0.1	0.0005	128
CIFAR-100	Shake-Shake (26 2x96d)	0.01	0.0025	128
CIFAR-100	PyramidNet+ShakeDrop	0.025	0.0005	64
R-CIFAR-10	Wide-ResNet-28-10	0.05	0.005	128
R-CIFAR-10	Shake-Shake (26 2x96d)	0.025	0.0025	128
SVHN	Wide-ResNet-40-2	0.05	0.005	128
SVHN	Wide-ResNet-28-10	0.005	0.001	128
SVHN	Shake-Shake (26 2x96d)	0.01	0.00015	128
R-SVHN	Wide-ResNet-28-10	0.05	0.01	128
R-SVHN	Shake-Shake (26 2x96d)	0.025	0.005	128



(a) Operation magnitudes.

(b) Normalized plot of operation probability parameters over time.

Figure 3. Plots showing the evolution of PBA operation parameters in a schedule learned on Reduced SVHN. Note that each operation actually appears in the parameter list twice; we take the mean parameter value for each operation in this visualization.